Precise Generation of Conformational Ensembles for Intrinsically Disordered Proteins *via* Fine-tuned Diffusion Models

Junjie Zhu[1,#], Zhengxin Li[1,#], Zhuoqi Zheng[1], Bo Zhang[1], Bozitao Zhong[1], Jie Bai[1], Xiaokun Hong[2], Taifeng Wang[3,*], Ting Wei[1,*], Jianyi Yang[4,*], Hai-Feng Chen[1,*]

[1]State Key Laboratory of Microbial Metabolism, Department of Bioinformatics and Biostatistics, SJTU-Yale Joint Center for Biostatistics, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 200240, Shanghai, China.

[2]College of Biological Science and Engineering, Fuzhou University, Fuzhou 350116, Fujian, China

[3]BioMap,10 Beilun Industrial Park, Yongteng North Road, 100080, Beijing, China.

[4]MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, 266237, Qingdao, China.

***Corresponding Author**
Email: haifengchen@sjtu.edu.cn; yangjy@sdu.edu.cn; weitinging@sjtu.edu.cn; taifeng@biomap.com

Notes

The authors declare that there is no conflict of interest.

Author Contributions

[#]These authors contributed equally to this work.

Abstract

Intrinsically disordered proteins (IDPs) play pivotal roles in various biological functions and are closely linked to many human diseases including cancer, diabetes and Alzheimer disease. Structural investigations of IDPs typically involve a combination of molecular dynamics (MD) simulations and experimental data to correct for intrinsic biases in simulation methods. However, these simulations are hindered by their high computational cost and a scarcity of experimental data, severely limiting their applicability. Despite the recent advancements in structure prediction for structured proteins, understanding the conformational properties of IDPs remains challenging partly due to the poor conservation of disordered protein sequences and limited experimental characterization. Here, we introduce IDPFold, a method capable of generating conformational ensembles for IDPs directly from their sequences using fine-tuned diffusion models. IDPFold bypasses the need for Multiple Sequence Alignments (MSA) or experimental data, achieving accurate predictions of ensemble properties across numerous IDPs. By sampling conformations at the backbone level, IDPFold provides more detailed structural features and more precise property estimation compared to other state-of-the-art methods. IDPFold is ready to be used in the elucidate the sequence-disorder-function paradigm of IDPs.

Introduction

Intrinsically disordered proteins (IDPs) constitute a category of proteins with unstable structures under physiological conditions[1]. These proteins, accounting for over 40% of eukaryotic proteomes[2,3], are involved in various biological functions including signal transduction, molecular recognition, and cell cycle regulation[4–6]. IDPs are also closely associated with various significant diseases, such as cancer, Parkinson's disease, and acquired immunodeficiency syndrome (AIDS)[7–9]. Unlike structured proteins that possess one or a few stable conformations, IDPs exhibit transitions between multiple conformations with very low energy barriers, constantly fluctuating within a broad ensemble of structures under physiological conditions[10–12]. Consequently, deciphering the conformational ensemble of IDPs poses a significant challenge for experimental methods such as X-ray, cryo-electron microscopy and NMR[13–16].

Currently, molecular dynamics (MD) simulation is the most commonly used and effective tool for sampling conformational ensembles[17,18]. By iteratively sampling target molecular system based on first principles, an estimation of conformational ensemble is obtained from simulation. However, MD simulation incurs high sampling costs, and ensuring thorough sampling of the ensemble is challenging. Additionally, commonly used force fields for IDPs, such as ff03CMAP[19], a99SB-disp[18], ESFF1[20], still exhibit considerable errors in estimating local and global properties of IDPs during simulations[21,22]. Although we can adjust MD simulations to better align with the characteristics of IDPs using experimental data, only a few of these proteins have been experimentally characterized[23]. These limitations constrain the application of MD in predicting the conformational ensemble of IDPs.

On the other hand, numerous deep learning methods have been developed and widely applied in the field of structure prediction for structured protein, such as AlphaFold2 and ESMFold[24,25]. Simultaneously, in protein design tasks, various generative models, such as RFdiffusion and Chroma, have been utilized for

generating diverse protein backbones[26,27]. Therefore, it is natural to consider whether deep learning methods can be employed for rapid and accurate prediction of conformational ensembles of IDPs. However, the structures associated with IDPs are currently very sparse. The PDB database contains $> 220{,}000$ structural entries[28], while the Protein Ensemble Database only includes 553 protein ensemble data[29]. Moreover, IDPs often lack high-quality multiple sequence alignment (MSA) data. MSA has been proven effective only in functional studies concerning folded or bound states of IDPs, and provides little assistance in predicting more disordered conformations[30]. These challenges make it highly daunting to use deep learning for robustly predicting the conformational ensembles of IDPs. Although Janson et al. have previously worked on IDP conformation generation and proposed idpGAN for predicting IDP conformational ensembles, they mainly forced on generating coarse-grained IDP conformations and often suffers from over-sampling, leaving a gap to coarse-grained or even all-atom MD simulations[31,32].

We introduce IDPFold here for predicting IDP dynamics directly from sequences based on a generative deep learning model. IDPFold utilized a protein language model to extract sequence information and further fed it into a structure generation module, enabling MSA-free conformation generation. To address the issue of insufficient data, we employed a hybrid dataset comprising crystal structures, NMR structures and MD trajectories for training IDPFold. The experimental structures enable the model to learn basic protein characteristics, while MD trajectories provide sufficient IDP structural data, ensuring accurate sampling on IDP systems. IDPFold generates IDP conformational ensembles at backbone level that is in better agreement with experimental observations than other state-of-the-art methods. IDPFold is able to sample conformational transitions from structured to disordered states, providing insights for studying the correlation between structures and functions of IDPs.

Results

IDPFold employs a conditional diffusion model framework for generating protein conformational ensembles from sequences (Figure 1a). This framework involves a forward diffusion process where noise is gradually added to real protein structures, and a reverse diffusion process where a deep learning network is used for denoising. By integrating specific protein sequence features into the model during the reverse diffusion process, we can generate conformational ensembles for specific proteins using this architecture. The denoising network takes as input the sequence features extracted by ESM2 and consists of an initialization block and four denoising blocks (Figure 1b). The initialization module integrates the sequence features, noise scale, and noised structure, while denoising modules combine Invarient Point Attention (IPA) with traditional Transformers to capture the chain-like structure within proteins and the rotational/translational state of each residue (Figure 1c). For more details, please refer to the Materials and Methods section.

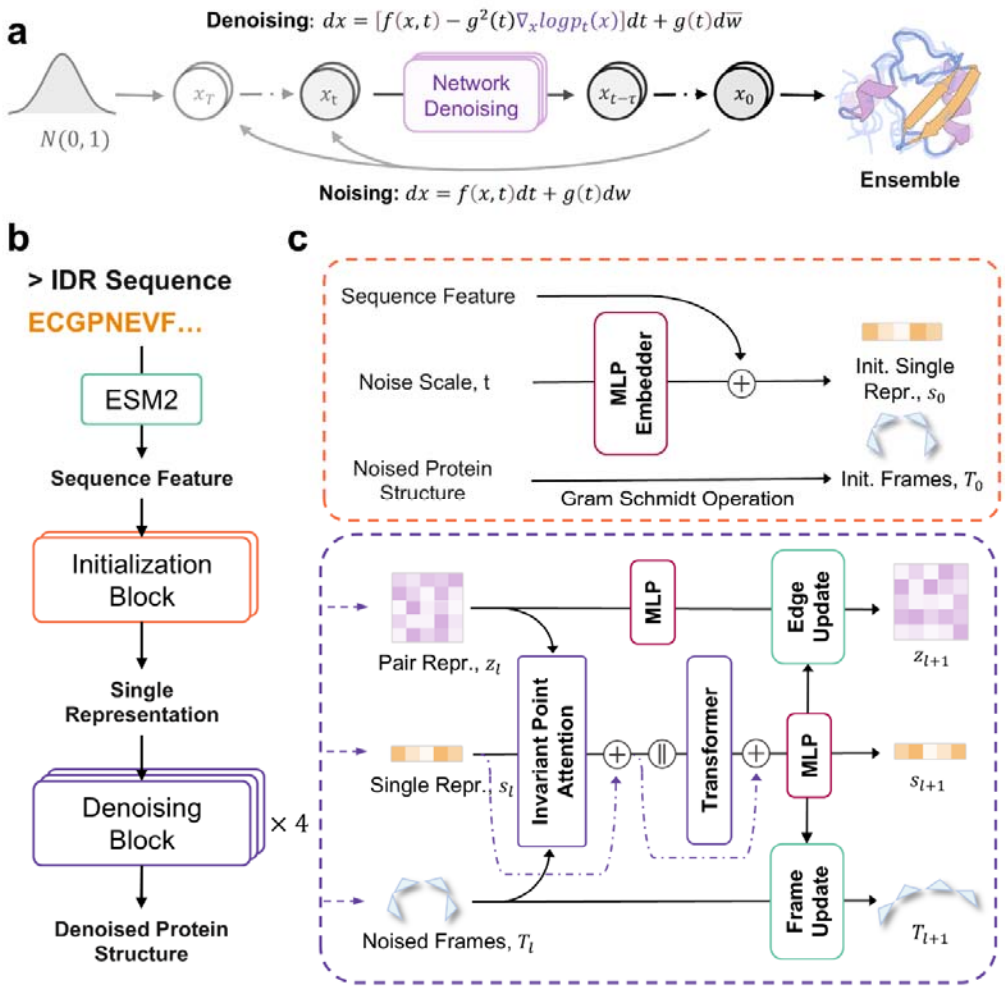

Figure 1: Architecture of IDPFold. **a,** Diffusion process for generating protein ensembles. **b,** Structure of denoising network. **c,** Detailed architecture of Initialization Block and Denoising Blocks.

IDPFold reproduces global features of IDPs

We first evaluated the IDPFold-predicted ensembles at coarse-grained level, primarily focusing on the global characteristic of the predicted ensembles, specifically the radius of gyration (Rg). Rg is a critical physical quantity that reflects the global characteristic of IDPs with its magnitude positively correlating with overall protein looseness, making it an empirical measure to describe and distinguish IDPs from structured proteins[33,34]. We first referred to the evaluation by Tesei et al. on the coarse-grained force field CALVADOS 2 by calculating the Rg error of the IDPFold-predicted ensembles on their test set. After removing sequences presented in the training set, this test set contained 58 IDP systems. Among these proteins, the average Rg error of the IDPFold-predicted ensembles was -3%, comparable to the performance of coarse-grained simulations reported by Tesei et al. (Figure 2a). We further examined the Rg distribution of predicted ensembles for each system and compared it with coarse-grained trajectories. The results demonstrated that the range of Rg distribution of IDPFold-predicted ensembles closely matched that of the coarse-grained simulations for most systems, indicating that IDPFold is capable of sampling a broad magnitude of IDP conformations similar to those observed in coarse-grained simulation trajectories and experimental values as well (Figure 2b, Figure S1).

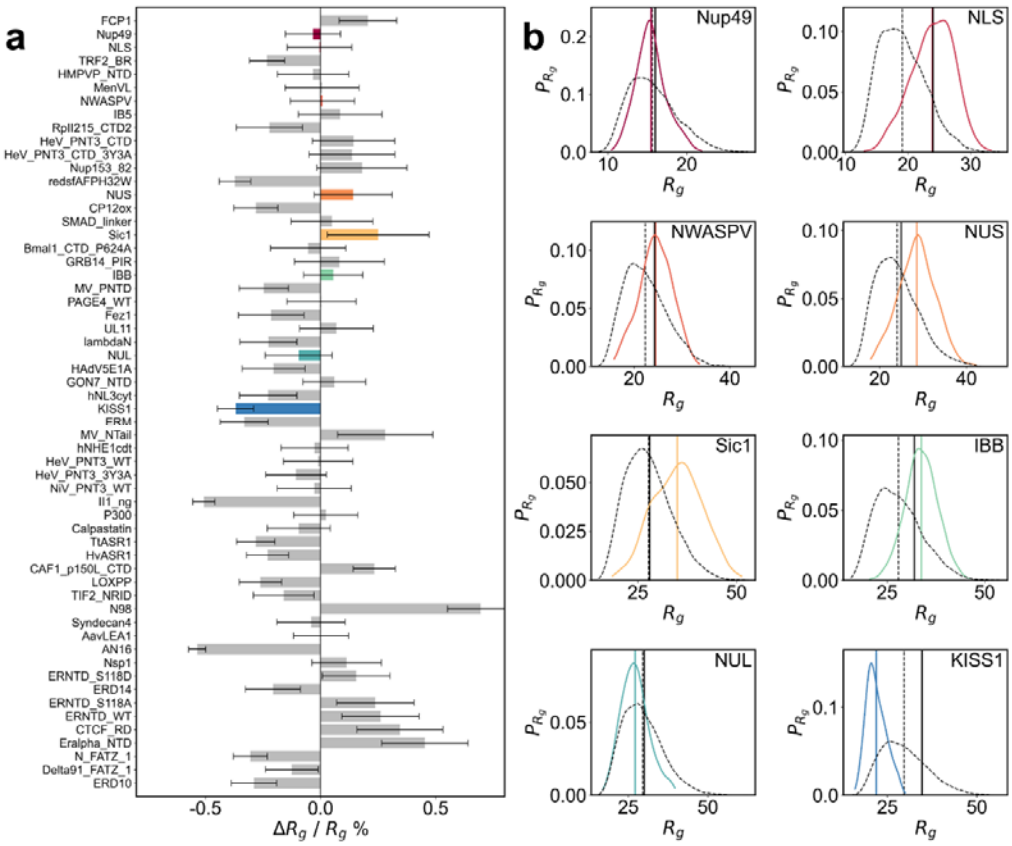

Figure 2: Comparison between the generated ensembles of IDPFold and coarse-grained MD simulation with CALVADOS 2. **a,** Relative error between IDPFold predicted and experimental radii of gyration on test set of CALVADOS 2. **b,** Rg distributions of IDPFold generated ensembles (colored) and CALVADOS simulations (black dashed). Experimental values are plotted as solid black lines.

IDPFold captures the global distribution of IDP ensembles

To further assess the robustness of IDPFold, we collected 27 IDP systems which possessed rich experimental observation data and were not present in the training set (Table S1, S2). We compared the IDPFold-predicted ensembles with the simulation results from CALVADOS 2 and the predictions from the coarse-grained deep learning method idpGAN on these systems[23,31]. The result showcased that average Rg error of

the IDPFold-predicted ensembles on this test set ($\varepsilon_{Rg} = -0.06$) was significantly smaller than that of idpGAN ($\varepsilon_{Rg} = -0.12$, with a paired t-test p-value of 0.02). The Rg error of the simulation trajectories was 0.02, which is better in absolute terms than both deep learning methods while without significant difference with IDPFold (a paired t-test p-value of 0.12). Compared to coarse-grained simulations, IDPFold tends to slightly underestimate Rg for IDPs, particularly for longer proteins. We hypothesize that this is partly due to the high conformational space complexity of long IDPs, making it more challenging to model their conformational ensembles. This is partially verified by the performance of the fine-tuned version of IDPFold, where the fine-tuning on MD trajectories corrected the disorder tendency in the generated conformations to a certain extent, while still underestimates the extent of extension for large systems. Overall, IDPFold tends to have a lower proportion of highly extended conformations in the predicted ensembles for large systems. This opens a future direction towards the improvements on the accuracy of global characteristic estimates by reweighting to increase the proportion of these extended conformations in the ensembles.

We next observed the ensemble distributions and main conformations generated by the three methods on specific cases. Here, we calculated the Rg-RMSD distribution of each ensemble using the initial conformations from coarse-grained simulations as a reference. The results showed that idpGAN exhibited over-sampling across all test systems and had significant deviations in estimating the main conformations. As a comparison, the ensemble distribution estimated by IDPFold was closer to coarse-grained trajectories, with its estimation of the free energy well positions more accurate than that of idpGAN (Figure 3a, c, e). Although the sampling range of IDPFold was smaller than that of coarse-grained trajectories, its estimation of the ensemble's Boltzmann distribution was more precise than that of idpGAN.

From the clustering results, the main conformations and their proportions sampled by IDPFold across the three systems were closer to the simulation results than idpGAN. In the Histatin5 protein, both IDPFold-predicted ensembles and the

simulation trajectories had an extended conformation as the main structure, whereas the main conformation predicted by idpGAN was in a semi-folded state (Figure 3b). In the two larger systems, Human Calpastatin and β Synuclein, the issues with idpGAN were even more pronounced, as it predicted collapsed main conformations, which led to a severe underestimation of the average Rg of these two proteins (Figure 3d, f). This indicates that idpGAN's estimation of ensemble distribution is inaccurate. On the other hand, IDPFold-predicted ensembles more accurately captured the shapes of the main conformations. Although IDPFold underestimated the Rg on β Synuclein due to overestimating the helical tendency in the structure, its estimates of the proportions of extended and folded conformations were closer to the simulation results.

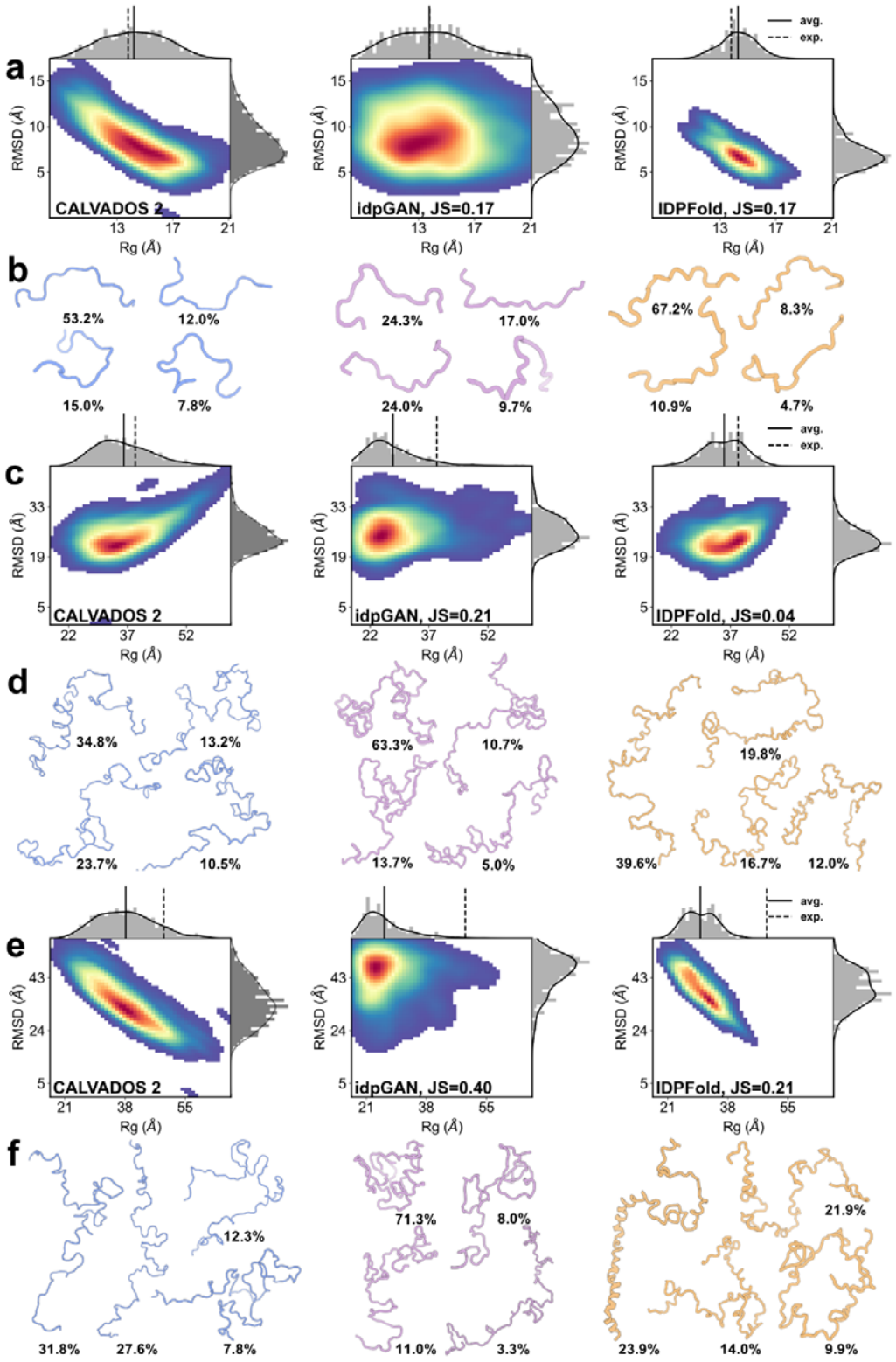

Figure 3: Comparing the generated ensembles of IDPFold and coarse-grained deep learning method idpGAN. **a, b,** Rg-RMSD distribution **(a)** and cluster centers **(b)** of

predicted ensembles on Histatin5. **c, d,** Rg-RMSD distribution **(c)** and cluster centers **(d)** of predicted ensembles on Human Calpastatin. **e, f,** Rg-RMSD distribution **(e)** and cluster centers **(f)** of predicted ensembles on β Synuclein.

IDPFold outperforms conventional all-atom simulation

A major advantage of IDPFold compared to the aforementioned coarse-grained simulations is its capability to generate protein backbones, allowing us to better understand the dynamic properties of proteins at higher precision. Thus, we conducted all-atom molecular dynamics (MD) simulations on 27 IDP systems aforementioned in the test set to assess the quality of the local features in the IDPFold-predicted ensembles. We used the ESFF1 force field which was specifically parameterized for IDPs, and solvent model OPC3-B[20,22]. To demonstrate the convergence of our simulation, we performed three sets of repeated simulations on 10 of these proteins, recording the distribution and mean of experimentally observed physical quantities across the parallel trajectories. The results of the convergence analysis are shown in Figures S2. Subsequently, we compared the IDPFold-predicted ensembles with the results from the all-atom MD simulations.

We first examined the distributions of bond lengths, bond angles, and dihedral angles predicted by IDPFold. In terms of bond length and bond angle distributions, the ensemble predicted by IDPFold closely resembles those observed in the MD trajectories, indicating that the model has effectively learned the arrangement of side chains in protein residues, accurately predicting the distances and arrangements between adjacent atoms (Figure 4a). Additionally, the distribution of Ω angles in the conformations generated by the model resembles that observed in MD trajectories showcased a peak around 180°, suggesting that the local peptide bond plane conformations generated by the model are stable and align with general protein characteristics (Figure 4b). Moreover, the Ramachandran plot of $\varphi - \psi$ angles shows

that the backbone dihedral angles of generated conformations predominantly fall within reasonable regions. The density estimation across various regions also closely approximate those observed in MD trajectories (Figure 4c-e). Specifically, with IDPFold fine-tuned on IDRome data, there is a notable improvement on the probability of the ppII region in the upper left corner compared to the untuned version, showing a distribution closer to MD trajectories. This indicates that the fine-tuning enables the model to generate more disordered structures, capturing the intrinsic biases of target proteins more precisely.

We next assessed ensemble average Rg of the generated ensembles (Figure 4f, g). We calculated the average error and standard deviation relative to experimental Rg values for conformational ensembles generated by IDPFold. Those generated by IDPFold pretrained on experimental structures significantly underestimated Rg for most IDPs (average error of -0.22), indicating a clear bias towards structured protein in prediction. This observation aligns with the high proportion of structured regions in experimental structures, lacking representation of natural disordered regions. However, the model significantly improved its Rg estimation after fine-tuning on MD trajectories, with an average error of only -0.06. Paired t-test demonstrates a significant improvement over all-atom simulation (average error of -0.19) with $p = 3.57 \times 10^{-4}$.

IDPFold not only accurately captured the ensemble average Rg of IDPs but also exhibited considerable diversity in generated conformations comparable to MD simulations (Figure S3, S4). Particularly, for IDP systems with a higher structured tendency, such as ACTR and drkN SH3, the pretrained IDPFold exhibited minimal variance and mean Rg in generated structures, indicating limited diversity and excessive compactness. As a comparison, the fine-tuned model showcased average Rg values closer to experimental observations with deviations comparable to MD trajectories ($p = 0.22$ in paired t-test), suggesting both a higher diversity in generated conformations and better alignment with experimental states.

Additionally, we evaluated the local features of generated ensembles using

chemical shifts as the target physical quantity to measure how generated ensembles differ from experimental observations (Figure 4A, Figure S5). Specifically, we computed C_α secondary chemical shifts of generated conformations and the Root Mean Square Deviation (RMSD) between ensemble average and experimental values. C_α secondary chemical shifts can reflect the secondary structure propensity of a protein system, where positive values indicate a higher helical propensity in certain region, negative values indicate a tendency to form strands, and values close to zero indicate disordered regions. The ensembles generated by fine-tuned IDPFold are close to experimental observations, showing an average RMSD of 0.53. Pairwise t-tests revealed significant improvement over MD simulations in fitting chemical shifts ($p = 0.03$), whereas pre-trained IDPFold performed worse than MD in fitting chemical shifts.

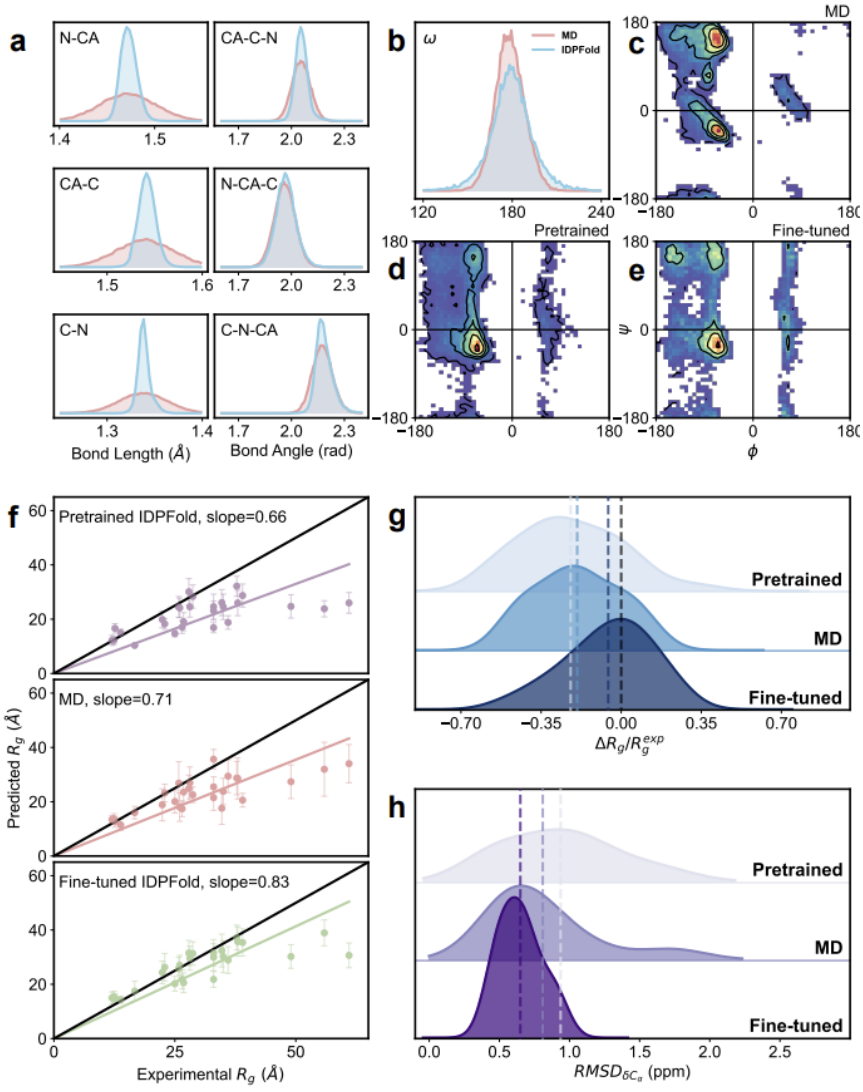

Figure 4: Comparison between the generated ensembles of IDPFold and all-atom MD simulation with ESFF1. **a-c,** Bond length **(a)**, bond angle **(b)** and omega angle **(c)** distributions of IDPFold generated ensembles (colored) and MD simulation (black). Lines in orange denotes inter-residue bond properties. **d-f,** Ramachandran plots of MD simulations **(d)**, pretrained IDPFold generated ensembles **(e)** and finetuned IDPFold generated ensembles **(f)**. **g, h,** Ensemble average Rg error **(g)**, RMSD on ensemble average C_α chemical shift **(h)** of MD trajectories (middle), pretrained (upper) and fine-tuned (lower) IDPFold against experimental values. **i,** Regression plot between experimental Rg values and predicted Rg values for MD simulation and IDPFold. The intercept was set to zero.

IDPFold predicts IDP ensembles comparable to long MD simulations

Additionally, we collected trajectories used by Robustelli et al. in testing the a99SB-disp force field, along with other force field trajectories for comparison[18]. We analyzed the differences in performance between IDPFold and four all-atom molecular dynamics simulations on seven proteins that overlapped with our test set. These 30-microsecond trajectories were confirmed to have converged. We primarily focused on the Rg error and the RMSD of C_α and C_β secondary chemical shifts (Figure 5a).

In terms of chemical shift accuracy, the error in the IDPFold-predicted ensembles reached a level comparable to a99SB-disp, surpassing traditional force fields such as a99SB-ILDN and CHARMM36m[17,35]. Furthermore, IDPFold demonstrated significantly better performance in Rg error compared to all four force fields, indicating that IDPFold accurately captures the global characteristics of IDPs while achieving a comparable level of precision in local structure prediction as current state-of-the-art force fields. Using α Synuclein protein as an example, we illustrated the differences between the ensembles sampled by IDPFold and MD trajectories from the a99SB-disp force field. We calculated the Rg-RMSD distributions from both methods and found that while a99SB-disp sampled a wider range of conformations, capturing more extended states, IDPFold predominantly sampled more compact conformations (Figure 5b). In terms of Cα chemical shifts, we observed that IDPFold significantly overestimated the helical propensity in the N-terminal region of α Synuclein, while achieving a similar level of accuracy as a99SB-disp in the C-terminal region (Figure 5c, d). This also explains the tendency of IDPFold to sample more compact structures. For C and HN secondary chemical shifts, RMSDs of IDPFold were slightly larger than those of a99SB-disp, while in $^3J(HN, H_\alpha)$ scalar coupling, IDPFold slightly outperformed a99SB-disp. Overall, we conclude that IDPFold's characterization of

IDP local features is on par with traditional force fields, and its sampling of α Synuclein takes around 20 minutes, significantly faster than the hundreds of hours required by traditional MD simulations. This demonstrates the power of IDPFold to serve as a complementary tool to traditional force field sampling which might pave the way for the exploration of macromolecular systems on a considerable time scale.

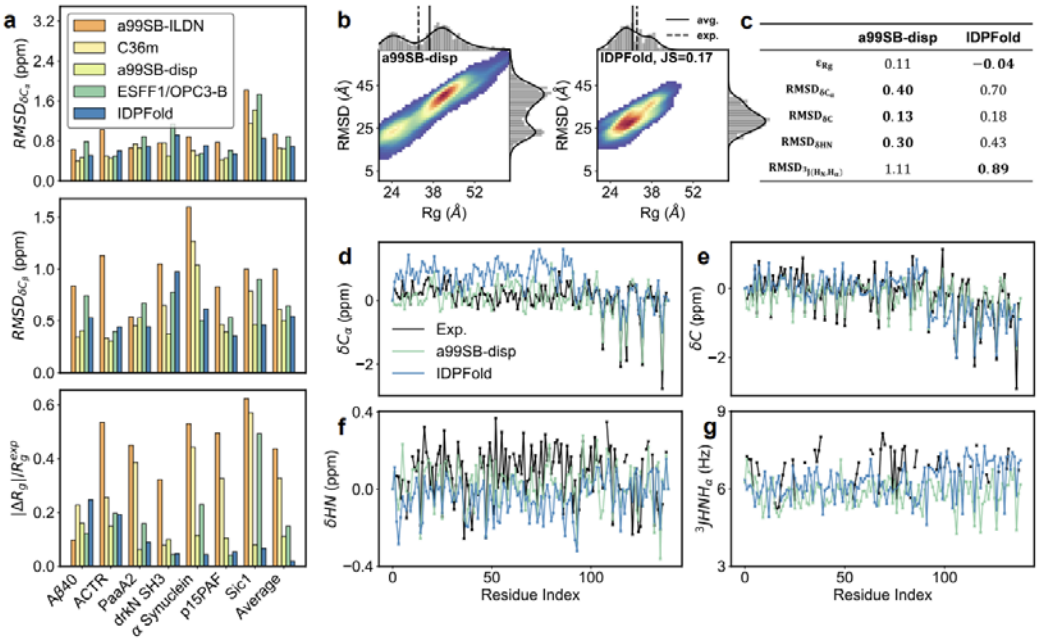

Figure 5: Comparison between the performance of IDPFold and that of all-atom MD simulations with different force fields. **a,** Accuracy of IDPFold in estimating local features (C_α, C_β chemical shifts) and global feature Rg compared to MD simulations. **b,** Rg-RMSD distribution of IDPFold generated ensemble compared to MD simulation with a99SB-disp on α Synuclein. **c,** Errors between generated ensembles and experimental observations for a99SB-disp and IDPFold on α Synuclein. **d-g,** Local Features including C_α chemical shifts **(d)**, C chemical shifts **(e)**, HN chemical shifts **(f)** and ${}^3J(HN, H_\alpha)$ **(g)** of IDPFold generated ensembles and MD simulation with a99SB-disp on α Synuclein.

IDPFold surpasses current state-of-the-art deep learning methods

Lastly, we compared IDPFold with existing deep learning methods on the test set, using experimental observations as the primary evaluation metric to establish a benchmark for ensemble prediction methods (Table 1, Figure S6). This benchmark included three coarse-grained methods and four methods with backbone-level or higher accuracy, in which AlphaFlow of the best performance (AF-PDB-base) is presented (Table S3). For local features, IDPFold achieved the best performance in most of the metrics among the backbone-precision methods (Figure S7-S13). In terms of global characteristic like Rg, IDPFold was slightly outperformed by the coarse-grained force field CALVADOS 2. We also evaluated the validity of the conformations generated by various deep learning methods (as defined in the materials and methods section). The results showed that all existing deep learning methods generated conformations with high validity. However, we observed a slight decrease in validity of IDPFold-generated ensembles after fine-tuning, likely due to the greater local structural fluctuations in the trajectory data used for fine-tuning compared to experimental structures which led to more pronounced local fluctuations in the generated conformations (Figure S14). Overall, IDPFold-predicted ensembles outperformed existing deep learning methods in all experimental observations and were comparable to or even better than traditional MD simulations.

Table 1. Benchmark on IDPFold and other methods. Bold values denote the best.

Methods	Validity	ε_{Rg}	$RMSD_{\delta C_\alpha}$	$RMSD_{\delta C_\beta}$	$RMSD_{^3J(H_N,H_\alpha)}$	$RMSD_{RDC_{NH}}$
CALVADOS 2[36]	-	**0.02**	-	-	-	-
idpGAN[31]	0.91	−0.12	-	-	-	-
idpSAM[37]	0.95	−0.55	-	-	-	-
MD 3*1μs ESFF1+OPC3-B	-	−0.19	0.81	0.64	**0.96**	3.37
AF-cluster[38]	**0.99**	−0.12	0.74	1.30	1.81	3.84
AlphaFlow[39]	0.97	−0.24	0.71	1.16	1.26	4.02
IDPFold	0.95	−0.06	**0.65**	**0.53**	1.01	**3.27**

Discussion

In this study, we developed IDPFold, a deep learning-based tool for predicting IDP dynamics conformation ensembles. IDPFold adopts a conditional diffusion model architecture to perform an end-to-end conformational generation, integrating protein language model for sequence feature extraction and DenoisingIPA module for conformation denoising. Through a two-stage training strategy on experimental data and MD trajectories of IDPs respectively, IDPFold efficiently and accurately samples IDP dynamics conformational ensembles. Furthermore, we established a benchmark of conformation sampling methods on 27 protein systems that contain IDRs. IDPFold precisely capture the overall compactness and local secondary structure features of IDPs, with predicted ensembles exhibiting features with values closer to experimental observations compared to both existing MD-based and deep-learning methods. Additionally, IDPFold can sample conformational changes from structured to disordered states in proteins, demonstrating a wide sampling range and high efficiency in ensemble sampling. This capability provides important insights for studying the conformational changes and functions of IDPs.

IDPFold achieves accuracy levels comparable to, or even higher than, traditional MD simulations in estimating IDP conformational ensembles with its sampling process not restricted by energy barriers. This is not only essential for studying IDP conformations but also holds tremendous promise for dynamic proteins like allosteric proteins and enzymes, which undergo large-scale conformational changes during functional processes. Although this work focused on training the model for IDPs and direct prediction performance on allosteric protein ensembles might not be optimal, the methodology would likely to be transferred on diverse categories of proteins, enabling more accurate estimations of conformational transitions in these functionally important proteins (Figure S15, Movie S1).

Although IDPFold demonstrates robustness and precision in predicting

conformational ensembles, its inference time is relatively long compared to some of the existed deep learning methods with an average time of about 20 minutes to sample an entire system. While this speed is much faster than conventional MD simulations, there is room for further improvement on model efficiency. To achieve higher sampling accuracy and more precisely protein backbone features, we employed a more complex network architecture and longer diffusion steps which might well sacrifice sampling efficiency. To enhance the inference efficiency of the model, optimizations can be explored such as reducing the number of diffusion steps in the diffusion model and optimizing transformer components by replacing them with architectures that are more efficient in terms of time and space. Additionally, in this study, we only trained and tested IDPFold on single-chain proteins and did not evaluate it on protein complexes or multi-chain proteins. In future work, we aim to extend the model to these areas to efficiently sample and interpret how molecules dynamically interact with each other.

Conformational ensemble prediction represents a significant frontier in protein structure research following static structure prediction. Accurate characterization of protein conformational ensembles can help us understand the dynamics of proteins and crucial conformational changes they undergo when binding to substrates or undergoing biochemical reactions under physiological conditions. Although deep learning-based tools are not yet widely adopted and most dynamic analyses still relying on MD simulations, our work suggest that properties predicted using deep learning methods might offer higher accuracy than MD simulations, potentially serving as complementary or even alternative approaches to MD simulations. However, current deep learning-based conformational ensemble prediction methods still face several challenges, such as inaccurate estimation of conformational free energies, less robustness compared to traditional methods, and the inability to capture temporal autocorrelation between conformations (i.e., inability to capture dynamic features).

Several studies have indicated that the single-point energy estimation of molecular

force fields can be used to train diffusion models, allowing models to learn more comprehensively about conformational space in cases where structural data are scarce[40,41]. This training approach, which utilizes force field energies rather than structural data, enables more accurate capture of conformational space distribution and greater robustness. However, this training strategy also makes the model sensitive to empirical parameters of the force field, demanding high precision in molecular force field accuracy. Additionally, there are works based on flow models or score-matching models aiming to fit the dynamic characteristics of MD trajectories, which might be a solution to the current inability of generative models to capture temporal autocorrelation between conformations[42].

Apart from the challenges of sparse training data and limited model representation capabilities, there is another significant issue in conformational ensemble prediction tasks: the scarcity and lack of uniformity in evaluation metrics. While we collected a set of experimental observations as a gold standard for this study, a vast number of protein systems lack experimental annotations, making it challenging to assess the generative performance of computational models on these systems. Therefore, we believe that constructing confidence metrics for conformational ensemble generation akin to pLDDT for protein structure prediction is also promising for the evaluation and improvement of protein conformational ensemble models.

Materials and Methods

Datasets

The data for training IDPFold consist of three components: high-quality crystal structures collected from the PDB dataset, conformations ensembles resolved by NMR and molecular dynamics (MD) simulation dataset of IDPs obtained through extensive back-mapping and energy minimization on coarse-grained trajectories.

For crystal structures, we referenced trRosetta and gathered a total of 15,051 X-

ray resolved structures with resolutions ≤ 2.5Å and sequence redundancy $\leq 30\%$[43]. These structures provide fundamental protein characteristics for the model, such as bond lengths, bond angle distributions and chirality of C_α. However, crystal structures typically depict stable conformations of proteins, i.e., structured conformations. Training the model solely on crystal structures would significantly underestimate the disorder tendency of IDPs and result in low diversity in conformation generation for individual protein sequences. Therefore, we additionally collected 12,339 NMR resolved protein conformational ensembles from the PDB. We filtered these NMR entries based on 30% sequence similarity both internally and against crystal structures, resulting in 539 systems comprising a total of 10,454 structures. By blending crystal structures with NMR ensembles, we obtained a combined total of 25,495 experimental structures for the initial phase of IDPFold training. These structures were well-defined and exhibited a higher structured tendency, ensuring that the model captures the local physical characteristics of proteins effectively.

For the IDP trajectory data, we obtained large-scale coarse-grained simulation data from IDRome that recorded most IDRs in human proteome[23]. These coarse-grained simulations accurately capture global features of IDPs (and IDRs), such as average radius of gyration (Rg) of conformational ensembles, but they only retain C_α atoms, whose resolutions fail to meet the requirement for model training. Therefore, we selected all systems with lengths exceeding 256 residues, totaling 3,880 systems. Choosing larger systems is primarily because these long disordered segments encompass most sequence and structural characteristics found in smaller systems, and training the model on larger systems can help enhance its generalization ability. For these systems, we first used pdbfixer to restore structures from coarse-grained to all-atom and then performed 100 steps of energy minimization with ff14SB[44] force field for all protein structures. A total of 77,600 optimized all-atom conformations was used for the second phase of IDPFold training, these conformations globally exhibit more disordered and can aid the model in learning distinctive conformational features of natural IDPs.

Furthermore, data used to assess the performance of IDPFold generation included 27 IDP systems that are fully described by experiments and are not present in the training set. We conducted $1\mu s$ all-atom MD simulations for these systems with the IDP-specific force field ESFF1 and solvent model OPC3-B[20,45]. These all-atom simulation trajectories described the dynamics and thermodynamic characteristics of IDPs, which are suitable for evaluating the quality of IDPFold-generated conformational ensembles.

Formulation of IDPFold

Diffusion Modeling on Protein Structure

To enable IDPFold to capture Boltzmann distribution of protein conformations at equilibrium state, we employ Score-Based Generative Modeling (SGM) to learn the probability distribution from protein structure data. SGM can be represented by a diffusion process $x_t \in \mathbb{R}^n$ defined by stochastic differential equation (SDE)[46]. The forward diffusion process is characterized by the following equation:

$$dx = f(x,t)dt + g(t)dw \ \#(1)$$

where $t \in [0,T]$ is a continuous index and $w \in \mathbb{R}^n$ is the standard Wiener process (a.k.a., Brownian motion). $f(x,t) \in \mathbb{R}^n$ is a vector-valued function called the drift coefficient, and $g(t) \in \mathbb{R}$ is a scalar function called the diffusion coefficient. Then, the corresponding backward diffusion process, or denoising process can also be defined by SDE[47]:

$$dx = [f(x,t) - g^2(t)\nabla_x log p_t(x)]dt + g(t)d\overline{w} \ \#(2)$$

where $d\overline{w}$ is a standard Wiener process as continuous time t flows backwards from T to 0, and dt is an infinitesimal negative time step. $x_{t=0}$ here represents ground truth data, or the protein conformations, while $x_{t=T}$ is sampled from Gaussian distribution. Therefore, by solving the backward SDE process, we can sample diverse protein conformations that obey Boltzmann distribution from a Gaussian distribution. In equation 2, each term except the score of x, $\nabla_x log p_t(x)$, is solvable. Therefore, we

only needed a score-matching network to fit $\nabla_x log p_t(x)$ at each time step to achieve the purpose of generating conformations.

Based on the above-mentioned standard formulation of SGM, we further required that the entire conformation generation process should be SE(3)-equivariant, i.e., diffusion process and network transformation were not sensitive to global rotation and translation of protein structures. SE(3)-equivariance can be described by the following equation:

$$\mathcal{F} \circ \rho(x) = \rho \circ \mathcal{F}(x) \#(3)$$

where $\mathcal{F}$ denotes data transformations like network prediction and diffusion process, ρ represents global rotation and translation. Typically, a protein conformation x is characterized by Cartesian coordinates $c_i \in \mathbb{R}^3, 1 \leq i \leq N$, where N denotes number of atoms. However, data transformations on Cartesian coordinates is computationally intensive and does not easily satisfy SE(3)-equivariance. Therefore, we adopted backbone frame parametrization, representing the protein conformation x as $T_j \coloneqq [R_j, v_j], 1 \leq j \leq n$, where n denotes the number of residues. Each backbone frame includes a 3×3 rotation matrix $R_j \in SO(3)$ and a translation vector $v_j \in \mathbb{R}^3$. A frame T_j can represent Euclidean transformation for each atom in residue j from local coordinates c_{local} to global coordinates c_{global} as $c_{global} = T_i \circ c_{local} \coloneqq R_i\, c_{local} + v_i$. Following the approach described in FrameDiff[41], we independently handle the rotation matrix R_j on the $SO(3)$ manifold and the translation vector v_j in $\mathbb{R}^3$ during the diffusion process, formulated as follows:

$$dT_t = \left[0, -\frac{1}{2}\beta(t)Pv_t\right]dt + \left[\sqrt{\frac{d}{dt}\sigma^2(t)}dw^{(SO(3))}, \sqrt{\beta(t)}Pdw^{(\mathbb{R}^3)}\right] \#(4)$$

Where $\beta(t)$ and $\sigma(t)$ control the scale of noise during the diffusion process, $w^{\mathcal{M}}$ denotes Brownian motion defined on manifold $\mathcal{M}$, and $P: \mathbb{R}^{3n} \to \mathbb{R}^{3n}$ is used for removing the center of mass. During the forward diffusion or noising process, the addition of noise on rotation matrices is determined by the noise kernel $p_{t|0}(R_t|R_0)$, which is obtained from an isotropic Gaussian distribution on the $SO(3)$ manifold.

This distribution is formulated as:

$$\mathcal{IG}_{SO(3)}(R_t; R_0, \sigma^2) = f(\omega_{t|0}) := \frac{1-\cos(\omega_{t|0})}{\pi} \sum_{l=0}^{\infty} (2l+1) e^{-l(l+1)\sigma^2 \frac{\sin((l+0.5)\omega_{t|0})}{\sin(0.5\omega_{t|0})}} \#(5)$$

Here, $\omega_{t|0} = Axis_angle(R_0^T R_t)$ is the axis-angle transformed representation of the composed rotation matrix $R_0^T R_t$. As for the translation vector, its noise addition process follows an Ornstein-Uhlenbeck process, also known as VP-SDE. The noise kernel for the translation vector is relatively straightforward, converging ultimately to $\mathcal{N}(0, I)$ as shown in the following equation.

$$p_{t|0}(v_t|v_0) = \mathcal{N}\left(v_t; v_0 e^{-\frac{1}{2}\int_0^t \beta(s)ds}, I - I e^{\int_0^t \beta(s)ds}\right) \#(6)$$

Network Design and Training

To achieve the goal of predicting conformational ensembles from sequence, we devised a sequence-conditioned score-matching model $s_\theta(x_t, x, seq)$ for denoising processes. As the diffusion process is designed to be strictly SE(3)-equivariant, network transformations should evidently preserve this property. Therefore, we adopted a variant of structure module from AlphaFold2 (Figure 1A) for updating backbone frames[24,32]. Here, we employed the Point Invariant Attention (IPA) mechanism to capture interactions and relationships between nearby residues, followed by a Transformer to learn global features and long-range interactions. This architecture has been demonstrated in previous research to promote training and generation of high-quality protein conformations. The aforementioned network design requires three inputs foreach layer: a one-dimensional vector representation s_l, pairwise feature representation z_l, and the set of rotation and translation updates T_l. We utilized ESM2-650M to extract protein sequence features, concatenated with residue position encoding and time encoding represented by trigonometric functions, to form the initial one-dimensional vector representation s_0. Pairwise feature representation z_0 is derived from s_0 based on relative positional encoding. After each layer of IPA-Transformer transformation, we updated the one-dimensional vector

representation through a fully connected network and subsequently updated the pairwise features *via* cross product.

The objective of score-matching networks differs from conventional neural network training goals. It does not aim to fit protein conformations directly but rather the scores of data perturbed to a certain degree, i.e., fitting the distribution of perturbed data. To measure how well the predicted scores fit actual distribution, we computed the DSM loss as follows:

$$\mathcal{L}_{dsm} = \mathbb{E}_{t\in[0,T]}\left\{\lambda(t)\mathbb{E}_{T_0}\mathbb{E}_{T_t|T_0}\left[\left|\left|s_\theta(T_t,t) - \nabla_{T_t} log p_{t|0}(T_t|T_0)\right|\right|^2\right]\right\} \#(7)$$

To ensure that the DSM loss at all time steps t results in a perfect fit score of 1, ensuring equal contribution of each time step to the loss function, we set the weights as follows:

$$\lambda(t) = \frac{1}{\mathbb{E}\left[\left\|\nabla_{T_t} log p_{t|0}(T_t|T_0)\right\|\right]} \#(8)$$

Additionally, to ensure the model learns detailed features of protein structures, apart from the DSM loss on rotation and translation matrices, we also incorporate mean square error (MSE) supervision for the positions of backbone atoms and differences on the distance matrix for samples with fewer forward diffusion steps ($t < \frac{T}{4}$). Therefore, the complete network training loss function can be represented as:

$$\mathcal{L} = \mathcal{L}_{dsm} + \omega_1\mathcal{L}_{bb} + \omega_2\mathcal{L}_{dist} \#(9)$$

where $\omega_1 = \omega_2 = 0.25$ control the weight of conformation quality loss.

During the training process, with a maximum time step $T = 1.0$, we optimize the DenoisingIPA using an Adam optimizer with the learning rate of 10^{-4}[48]. For the translation vector part of the network training, we employ a linear noise strategy within the VP-SDE framework, while for the rotation matrix part, we use a logarithmic noise strategy within the VE-SDE framework, as shown in the following equation[46]:

$$\beta(t) = \beta_{min} + \frac{t}{T}(\beta_{max} - \beta_{min}),$$
$$\beta_{min} = 0.1, \beta_{max} = 20 \#(10)$$

$$\sigma(t) = \log(te^{\sigma_{max}} + (T - t)e^{\sigma_{min}}),$$
$$\sigma_{min} = 0.1, \sigma_{max} = 1.5 \#(11)$$

Implementation Details

The training of IDPFold consists of two stages: pre-training on experimental structures and fine-tuning on MD trajectories. In Table 2, we present the model hyperparameters utilized during the 2 training stages:

Table 2. Hyperparameters and training details of IDPFold

Hyperparameters	**Training Stage**	
	Training on Experimental Data	Training on MD Trajectories
Single Repr. Channel	256	
Pair Repr. Channel	128	
Hidden Channel	256	
IPA Layers	4	
Transformer Layers	2	
Transformer Heads	8	
Model Size	17.8 M parameters	
Learning Rate	10^{-4}	10^{-5}
Batch Size	8	32
Iterations	1.12 M	0.44 M
Time	~9 GPU days	~15 GPU days

Evaluation Metrics and Analysis Tools

We evaluated the quality of IDPFold-predicted IDP conformational ensembles from two perspectives: local features of the generated structures and global features of the conformational ensembles. For local features, biotite is used to compute inter-residue bond lengths and bond angles of all generated conformations[49]. We also

analyzed the distribution of backbone dihedral angles to assess the model's prediction accuracy regarding C_α chirality and secondary structure. Additionally, we used mdtraj to calculate scalar coupling between HN and H_α, employed SPARTA+ to compute chemical shifts, and then performed ensemble averaging[50,51]. These analyses helped determine whether the local environment of the protein backbone aligns with experimental observations.

For global features, we calculated Rg of the generated conformations and compared the ensemble average with experiment results. Additionally, we calculated RMSD of generated conformations against the initial structures utilized in MD simulation in order to construct Rg-RMSD space. Through clustering the generated conformations and projecting them onto Rg-RMSD space, we aim to discern the diversity of generated conformations and assess the model's learning efficacy in capturing the Boltzmann distribution information from MD trajectories in the training dataset. MMTSB toolset is applied for conformation clustering.

Due to the fact that most current work on generating protein conformational ensembles from sequences adopts a coarse-grained representation, where only C_α coordinates are generated, we first fixed the generated structures with pdbfixer and ran a 100-step energy minimization. During minimization we added restraints on C_α to make sure the optimized backbones did not differ from the original ones too much. This process cost about 6 seconds per conformation. Then we evaluated Validity and Fidelity and compared the performance of IDPFold with previous methods.

Validity assesses whether the generated conformations contain unreasonable C_α distances. In a protein structure, due to van der Waals interactions between atoms, C_α distances should not be too close. Additionally, because neighboring C_α in the protein backbone are connected by $C - N$ with specific bond lengths, C_α distances should not be excessively far apart. Therefore, following the approach of Str2Str, we defined a reasonable range for C_α distances as follows:

$$\delta_{vdw} < d_{C_\alpha^i, C_\alpha^{i+1}} < \delta_{bond} \#(12)$$

where $\delta_{vdw} = 2 \times 1.7 - 0.4$ is defined as the sum of two C_α van der Waals radius

minus an acceptable overlap distance of 0.4[52]. δ_{bond} is taken as the maximum C_α distance observed in MD trajectory of each test system. Since a reasonable range is defined, we could define validity as proportion of valid conformations. A higher validity indicates a lower probability of mis-estimated C_α distances, thus demonstrating better performance.

Fidelity refers to how well the model-generated ensembles match experimental observations. We selected Rg, C_α and C_β secondary chemical shifts, J-coupling constants between HN and H_α, and backbone N-HN RDCs as target physical quantities to measure the fidelity of models, calculating the errors between model-generated conformational ensembles and experimental observations using the following formula:

$$\varepsilon_{Rg} = \frac{\langle Rg \rangle^{pred} - \langle Rg \rangle^{exp}}{\langle Rg \rangle^{exp}} \#(13)$$

$$RMSD_{\langle A \rangle} = \sqrt{\mathbb{E}(\langle A \rangle^{pred} - \langle A \rangle^{exp})^2}$$

where $\langle Rg \rangle$ denotes the ensemble average Rg, $\langle Rg \rangle^{pred}$ is the generated ensemble average Rg while $\langle Rg \rangle^{exp}$ is experimental observation. Similarly, $\langle A \rangle^{pred}$ and $\langle A \rangle^{exp}$ denotes ensemble average physical quantity (e.g., chemical shifts and J-couplings) and experimental observation respectively. Using the above definitions, we calculated ε_{Rg}, $RMSD_{\delta C_\alpha}$, $RMSD_{\delta C_\beta}$, $RMSD_{^3J(H_N,H_\alpha)}$ and $RMSD_{RDC_{NH}}$ for benchmarking current conformation generation methods, aiming for values close to zero for all. ε_{Rg} reflects the reasonableness of the model's generated conformations in terms of their compactness, while $RMSD$s on other physical quantities indicate how well the generated conformations conform to experimental observations on local structures.

For evaluating the conformational ensembles generated by the models, all methods except AF-cluster generated 300 conformations for subsequent evaluation[53,54]. This number of conformations has been previously shown to adequately reflect the structural diversity for intrinsically disordered proteins of

similar size to the largest protein in our test set. Additionally, we conducted convergence test on IDPFold-generated ensembles as depicted in Figure S16. The number of conformations generated by AF-cluster is influenced by the number of clusters in the MSA, and we used its default MSA and clustering settings. All conformation generation was performed on NVIDIA A100 GPUs.

Acknowledgements

This work was supported by the Center for HPC at Shanghai Jiao Tong University, and the National Key Research and Development Program of China (2020YFA0907700 and 2023YFF1205102), the Fundamental Research Funds for the Central Universities (YG2023LC03), and the National Natural Science Foundation of China (21977068, 32171242 and T2222012).

Code Availability

The code of IDPFold is available at https://github.com/Junjie-Zhu/IDPFold.

References

1. van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
2. Dunker, A. K. *et al.* The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9**, S1 (2008).
3. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **25**, 187–211 (2024).
4. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **41**, 6573–6582 (2002).
5. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins in Human

Diseases: Introducing the D^2 Concept. *Annu. Rev. Biophys.* **37**, 215–246 (2008).
6. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *J. Mol. Biol.* **323**, 573–584 (2002).
7. Zhao, R. *et al.* Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev.* **14**, 981–993 (2000).
8. Singleton, A. B. *et al.* α-Synuclein Locus Triplication Causes Parkinson's Disease. *Science* **302**, 841–841 (2003).
9. Casu, F., Duggan, B. M. & Hennig, M. The Arginine-Rich RNA-Binding Motif of HIV-1 Rev Is Intrinsically Disordered and Folds upon RRE Binding. *Biophys. J.* **105**, 1004–1017 (2013).
10. Fernández, A. & Scheraga, H. A. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 113–118 (2003).
11. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
12. Mittag, T. & Pappu, R. V. A conceptual framework for understanding phase separation and addressing open questions and challenges. *Mol. Cell* **82**, 2201–2214 (2022).
13. Chruszcz, M., Borek, D., Domagalski, M., Otwinowski, Z. & Minor, W. X-ray diffraction experiment – the last experiment in the structure elucidation process. *Adv. Protein Chem. Struct. Biol.* **77**, 23–40 (2009).
14. Dandey, V. P. *et al.* Time-resolved cryoEM using Spotiton. *Nat. Methods* **17**, 897–900 (2020).
15. McDermott, A. Structure and Dynamics of Membrane Proteins by Magic Angle Spinning Solid-State NMR. *Annu. Rev. Biophys.* **38**, 385–403 (2009).
16. Jensen, M. R., Zweckstetter, M., Huang, J. & Blackledge, M. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chem. Rev.* **114**, 6632–6660 (2014).
17. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
18. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758–E4766 (2018).
19. Zhang, Y., Liu, H., Yang, S., Luo, R. & Chen, H.-F. A Well-balanced Force Field ff03CMAP for Folded and Disordered Proteins. *J. Chem. Theory Comput.* **15**, 6769–6780 (2019).
20. Song, D., Liu, H., Luo, R. & Chen, H.-F. Environment-Specific Force Field for Intrinsically Disordered and Ordered Proteins. *J. Chem. Inf. Model.* **60**, 2257–2267 (2020).
21. Zhong, B., Song, G. & Chen, H.-F. Balanced Force Field ff03CMAP Improving the Dynamics Conformation Sampling of Phosphorylation Site. *Int. J. Mol. Sci.* **23**, 11285 (2022).
22. Mu, J., Pan, Z. & Chen, H.-F. Balanced Solvent Model for Intrinsically Disordered and Ordered Proteins. *J. Chem. Inf. Model.* **61**, 5141–5151 (2021).
23. Tesei, G. *et al.* Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024).
24. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
25. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
26. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

27. Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
28. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
29. Lazar, T. *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.* **49**, D404–D411 (2020).
30. Riley, A. C., Ashlock, D. A. & Graether, S. P. The difficulty of aligning intrinsically disordered protein sequences as assessed by conservation and phylogeny. *PLOS ONE* **18**, e0288388 (2023).
31. Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **14**, 774 (2023).
32. Lu, J., Zhong, B., Zhang, Z. & Tang, J. Str2Str: A Score-based Framework for Zero-shot Protein Conformation Sampling. in (2023).
33. Cragnell, C., Durand, D., Cabane, B. & Skepö, M. Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS. *Proteins Struct. Funct. Bioinforma.* **84**, 777–791 (2016).
34. Moncoq, K. *et al.* SAXS Study of the PIR Domain from the Grb14 Molecular Adaptor: A Natively Unfolded Protein with a Transient Structure Primer? *Biophys. J.* **87**, 4056–4064 (2004).
35. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).
36. Tesei, G. & Lindorff-Larsen, K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res. Eur.* **2**, 94 (2023).
37. Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. 2024.02.08.579522 Preprint at https://doi.org/10.1101/2024.02.08.579522 (2024).
38. Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
39. Jing, B., Berger, B. & Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles. Preprint at https://doi.org/10.48550/arXiv.2402.04845 (2024).
40. Zheng, S. *et al.* Towards Predicting Equilibrium Distributions for Molecular Systems with Deep Learning. Preprint at https://doi.org/10.48550/arXiv.2306.05445 (2023).
41. Yim, J. *et al.* SE(3) diffusion model with application to protein backbone generation. Preprint at http://arxiv.org/abs/2302.02277 (2023).
42. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
43. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* **117**, 1496–1503 (2020).
44. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
45. Pan, Z., Mu, J. & Chen, H.-F. Balanced Three-Point Water Model OPC3-B for Intrinsically Disordered and Ordered Proteins. *J. Chem. Theory Comput.* **19**, 4837–4850 (2023).
46. Song, Y. *et al.* Score-Based Generative Modeling through Stochastic Differential Equations. Preprint at http://arxiv.org/abs/2011.13456 (2021).
47. Anderson, B. D. O. Reverse-time diffusion equation models. *Stoch. Process. Their Appl.* **12**,

313–326 (1982).
48. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
49. Kunzmann, P. & Hamacher, K. Biotite: a unifying open source computational biology framework in Python. *BMC Bioinformatics* **19**, 346 (2018).
50. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
51. Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48**, 13–22 (2010).
52. Ramachandran, S., Kota, P., Ding, F. & Dokholyan, N. V. Automated Minimization of Steric Clashes in Protein Structures. *Proteins* **79**, 261–270 (2011).
53. Daughdrill, G. W. *et al.* Understanding the Structural Ensembles of a Highly Extended Disordered Protein. *Mol. Biosyst.* **8**, 308–319 (2012).
54. Schwalbe, M. *et al.* Predictive Atomic Resolution Descriptions of Intrinsically Disordered hTau40 and α-Synuclein in Solution from NMR and Small Angle Scattering. *Structure* **22**, 238–249 (2014).